DOCUMENT RESUME

ED 298 913 IR 013 298

AUTHOR Baker, Eva L.; Alkin, Marvin C.
TITLE Formative Evaluation of Instructional Development.
PUB DATE 84
NOTE 15p.; In: Instructional Development: The State of the Art, II; see IR 013 273. For volume I, see ED 292 444.
PUB TYPE Guides - Non-Classroom Use (055) -- Information Analyses (070) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Collection; *Evaluation Methods; Evaluation Problems; *Evaluation Utilization; *Formative Evaluation; *Instructional Development; *Models; Research Design

ABSTRACT

This chapter discusses current methods of formative evaluation and suggests that these methods themselves need to be subjected to more critical research and evaluation in order to provide them with a more scientific base. Topics addressed include problems and methods associated with data gathering, evaluation designs, number of subjects available, use of findings, formative evaluation models, decision making algorithms, expert judgments, and the use of data for guiding and improving various stages of the instructional and product development process. (70 references) (Author/EW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED298913

Chapter 25

Formative Evaluation of Instructional Development

Eva L. Baker
Marvin C. Alkin

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Charles Dills

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Research in the area of formative evaluation is presented in this chapter. The inadequacy of current formative evaluation procedures is pointed out. The chapter indicates that we are currently in an era of seers and clairvoyants in terms of formative evaluation, and provides suggestions which might lead to the evolution of formative evaluation into the domain of scientific application.

Substantial amounts of money are wasted each year on the purchase and installation of educational products which later prove to be inappropriate or ineffective. To prevent such economic and educational waste, and the negative effects it could have upon the future acceptance and use of educational products, and at the same time improve the products ultimately produced, developers should engage to a greater extent in formative evaluation of all products. These formative evaluation data would provide information to developers which would allow them to modify and improve their products before they are distributed.

The inadequacy of current formative evaluation procedures in the development of products has been well documented by Komoski (1971). He notes that less than 1% of the approximately 14,000 textbooks being sold are "systematically shaped through the learner try-out and revision process"--utilizing formative evaluation data for product modification. Komoski states that the situation is worse in the field of the educational film. In the area of broadcast video-taped instruction "only a little over 1% of the television materials used in schools has been learner verified" (Komoski, 1971).

Even in the area of programmed instruction, the area in which the basic techniques for learner verification through formative evaluation and revision were developed, the situation is discouraging. "EPIE's examination of 633 of the programmed items now in use in major curriculum areas in schools revealed that research evidence was available for only 7% of these materials while

* The authors acknowledge with appreciation the assistance of Ms. Laura Spooner and Mrs. Bea Kass.

The project reported herein is performed pursuant to a grant from the National Institute of Education, Department of Health, Education and Welfare and opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education; no official endorsement by the National Institute of Education should be inferred.

IR013298

BEST COPY AVAILABLE

some field testing was claimed for another 8%" (Komoski, 1971).

Deficiencies in the use of formative evaluation procedures as a part of instructional product development are not new and have been known for years. (Komoski, 1956). Apparently there is great willingness on the part of publishers of instructional materials of all types to turn their heads, make believe the malaise does not exist, and hope that, like the mumps, it will pass with time. Now, let us consider what is meant by the term "formative evaluation."

FORMATIVE EVALUATION

Other publications (Alkin, 1969), have referred to the purpose of evaluation as the provision of information for decision-makers and have maintained that "evaluation must take into consideration the ultimate decisions to be served." One classification of ultimate decisions was suggested by Scriven (1967), who made a distinction between formative and summative evaluation. Generally, formative evaluation is used to describe the evaluation of educational programs while they are still in some stage of development. Schutz (1970) has said, "the product of formative evaluation activities is expected to be an improved instructional program." Or, to state the definition of formative evaluation somewhat differently, formative evaluation is intended to provide data for instructional product revision, as opposed to summative evaluation, which is to provide data for product validation.

This distinction causes us some concern. Technically, we cannot conceive of a summative evaluation taking place with respect to instructional product development if, as is frequently the usage of the term, one considers that summative evaluation refers to the last major evaluation effort related to a product prior to its public release. Let us assume, then, that the intent of summative evaluation is to validate the materials. Can we say that, following this validation, not an idea, not a paragraph, not a phrase, not a word will be changed, and that the published product will be identical in every respect to the "validated" product? The strict definitions of formative and summative, therefore, seem to preclude the possibility of a summative evaluation taking place.

In this regard, we like the Stake (1967) distinction between formative evaluation and summative evaluation. Stake notes that "It is probably more useful to distinguish between evaluation **oriented to** (authors' italics) developer-author-publisher criteria, and standards and evaluation **oriented to** (authors' italics) consumer-administrator-teacher criteria and standards." Thus, with respect to the evaluation of instructional products, we may make the distinction between formative evaluation, evaluation **oriented toward** developers in order to provide them with information to be used in modifying and improving programs.

From this definition, it is obvious that a large part of the activity of product development deals with formative evaluation. Indeed, the role of the evaluator is generally not unique nor separate from his role in performing the activities of product development. In fact, formative evaluation is an inherent part of the product developer's role. Because product development and formative evaluation go hand in hand, they are frequently performed by the same people. In many cases, the supplier of this information, the evaluator, is also the developer. That is, the same individual has a dual role in the product-development process. As a result, the evaluator selects and supervises the product's field-testing plan, and translates the information he gets from the tests into product changes.

Thus, it is important to note that when we refer to "the formative evaluator" in product development, or to formative evaluation data, we are referring to an evaluative function or role that could be performed by the developer himself. This overlap of function may cause some difficulty in separating "research on formative evaluation" from "research (more generally) in product development." The next section will deal with this issue.

RESEARCH ON FORMATIVE EVALUATION

Difficulties in Doing Research

Attempts to apply formative-evaluation procedures greatly outweigh any systematic research investigations into the critical factors for conducting formative-evaluation work. This lack of experimental research into the process of evaluation should surprise no one. Research on "research" and research on "development" have been sparsely-supported activities, and research and development enterprises have longer histories and wider practice than the evaluation activities.

In order to conduct research on evaluation (or on research or development), a combination of ideal conditions in a real-life project must exist. There must be investigators who think that knowledge about the process of evaluation, for example, will be important or useful. There must be individuals who can and will defer their single-minded concerns for their mission objectives in order to design studies which more leisurely compare systematic variations in process. There must be support that permits the conduct of parallel treatments over a length of time sufficient to establish differences. Resources of such a magnitude are not often available.

In addition, a number of methodological problems impede such work. Methodology, research on evaluation might require, for example, in the simplest case, comparisons of evaluations that systematically include procedure S with evaluations that exclude it. As sufficient number of the evaluators must conform to each treatment varia-

tion, so that in effect of the variation, rather that the evaluators's finesse, is tested in the research. Given the wide variety of different educational programs and systems, and the usual idiosyncratic behavior of those called upon to be evaluators of programs, it is not very surprising that little real-life research has been done on evaluation.

The alternative to research conducted in the course of real projects is to use simulated settings for research studies. However, the use of simulated settings requires the investigator to make serious sacrifices in verisimilitude. For instance, graduate students are likely to be the group most readily available to serve as research subjects in simulated settings, for the possibility of obtaining "real-world" evaluators as participants in such experiments is usually remoted. There are several reasons why bias might be introduced through the selection of students as research subjects. One reason is that student-subjects are more likely to conform to treatment requirements than are practicing evaluators, who might not agree with the particular treatment variation to which they would be exposed. Moreover, student-subjects also lack real-world experience in evaluation--they also are usually well-protected from the administrative-personnel concerns that plague many on-the-line evaluators.

Another problem relative to the use of simulated settings is that the scope of the problems provided for a student to solve necessarily is often much narrower than is the scope of the problems confronting a project evaluator. The nature of the simulated setting requires the selection of a small number of manageable and easily manipulated dimensions thought to be indicative of the major aspects of the real-world setting. Studies have been conducted, however, in the area of formative evaluation, and have been subject to almost all of the deficiencies noted above. Since the issues addressed in this research have high potential payoffs, it may be far better to endure the possible deficiencies of the data collection strategies than not to conduct the research at all. The conduct of such studies may be viewed as valuable despite the caveats that should accompany the findings.

Most of the research on evaluation then, is subject to the limitation of having been conducted in contrived and usually simulated settings. The major areas of investigation of such studies can be categorized as primarily dealing either with (1) the manner in which data are gathered or (2) the use of formative evaluation data.

If one agrees that the task of the formative evaluator is to monitor programs in order to provide evaluation data leading, hopefully, to improved results, then it is not surprising that the focus of most research on formative evaluation has been at the data acquisition/utilization conjoint. The two major research questions related to formative evaluation are: (1) What are the major types of data to obtain and the major sources of such data? and (2) What is the impact of evaluation data upon the modification or programs?

Studies of Data Gathering

In order to assess the effectiveness of instructional systems, a variety of data types may be investigated. We will discuss those data types most frequently used in research efforts. At a subsequent point in this chapter, alternative data types for formative evaluation will be suggested. Among the most prominent information-bases for instructional improvement is the learners' criterion test performance. The extent to which learners have mastered the objectives of a given instructional sequence is one critical ingredient in the judgment of program effectiveness. Effects of instructional programs have typically been measured by objectives-based instruments such as those advocated popularly by Mager (1962) and subsequently articulated by Glaser (1963). These measures typically have been oriented toward the achievement of cognitive goals.

A coordinate and very popular source of data utilized in the formative evaluation of instructional programs is the within-program response, assessed by recording students' answers to problems they encounter enroute to the desired goals. Another important source of data are learners' attitudes. These attitudes are of several types: (1) toward the subject-matter itself, that is, how much they like mathematics; and (2) toward the instructional process, that is, how interesting they found the sequence. These are each significant sources of post-instructional effects. Attitudinal data may also be obtained by recording positive and negative responses made in the course of the instructional sequence.

Another important category of formative evaluation data we refer to as implementation data (Alkin, 1969); this relates to whether the instructional system under development is capable of being utilized in the specified way. The more complex the instructional system, the greater the need to ascertain its implementability. For example, an important implementation-evaluation finding would be evidence that an instructional system requires combinations and sequences of ancillary support materials that are either not likely to be available or are not likely to be mobilized and synchronized in the manner appropriate.

Within-Program/Posttest Performance

Although there have been many recommendations regarding the sorts of data to collect in program development (see Baker, 1973; Markle, 1967), research on data gathering has centered upon the consideration of cognitive achievement. One area of high concern among researchers has been the relationship of within-program responses to posttest performance. In part, this concern may be attributable to the fact that data was readily obtainable and quantifiable. When one recognizes that early research was in the field of programmed instruction, just once removed from operant conditioning, it is easy to understand the research

emphasis upon within-program responses.

The response rates, or error counts, in a learning sequence were thought by some to be a direct indication of the ease with which learning occurred, provided, of course, that the responses made during the program were congruent with the intended outcomes of the sequence. Theorists in programmed instruction differed regarding the intended function of errors in a learning sequence. Glaser (1966) suggested that within-program errors represented an inadequacy of the program itself; Crowder (1960) saw errors performing a diagnostic function, indicating what particular discriminations needed additional attention.

A series of research studies attempted to determine the relationship between response rates and criterion performance. In three separate research investigations, the relationship between errors and criterion performance was not found to be reliable. Silberman, Coulson, Melaragno, and Newmark (1964) manipulated errors without affecting criterion performance. Anderson and Faust (1967), in a study involving programmed instruction, observed the relationship of errors to the manner in which the questions were presented, and suggested that errors were not necessarily related to criterion behavior. Baker, Schulz, and Sullivan (1967) conducted a study in which low error rates (high numbers of correct responses) did not necessarily predict strong posttest performance.

From the present vantage point, one is not at all surprised at the findings of Silberman's, Anderson's, and Baker's studies. The studies were conducted on materials which were "programmed" according to the standards of the day. They represented a single development team's notions of sequence and design. At that time, little work had been conducted in assuring that both practice and criterion examples were drawn from the same item population. Thus, it was very easy to design simple practice items, items which were often susceptible to sequence prompts, and thus were not necessarily at the same level of difficulty as the criterion items. Further, the notion of domain-referenced testing, as described by Hively, Patterson, and Page (1968), Hively, et al. (1973), Baker (1973), and Millman (1973), had not surfaced in the formative-evaluation literature. Test design was enmeshed in the rituals of norm-referenced item construction and item improvement. Idiosyncracies in item design prevailed. So the lack of congruency of performance on within-program problems and criterion exercises was to be expected.

Posttest Data

Criterion test performance has nonetheless served as a basis for inferring program success or revision requirements. Unfortunately, little research has been conducted regarding the form and frequency with which such posttest data should be collected. An advance in program assessment strategy has been the work in matrix sampling. Husek and Sirotnik (1968) have described a procedure for reducing the amount of testing time requisite for program revision. Samples of test items are administered to samples of learners, and the sum across all items is used as an index of program success. Shoemaker (1972) has prepared a refined and expanded paper on the topic, describing multiple-matrix sampling, where samples of item-person matrices are repeated across time. Shoemaker demonstrates that such procedures are superior to usual non-sampled test administrations in terms of the obtained standard error of estimate, and no worse than regular procedures in terms of the bias of the data. Such procedures are particularly appropriate when student time is at a premium and a set of coordinate objectives may be assessed by the developer.

Fortunately, multiple-matrix sampling procedures are useful even when task hierarchies, or objectives requiring a sequence of particular subskills, are to be assessed. The item sample must be stratified so as to provide each learner with an entire hierarchy, suppose, of a terminal objective and two enabling or subordinate objectives. Without stratification, matrix sampling would be of little use for the prescriptive revision of the instructional sequence under scrutiny. Gagne and Paradise (1961) conducted research that generated hypotheses to explain unachieved criterion objectives. Using a hierarchical explanation of learning, they suggested that failure in criterion performance may be attributable to (1) the omission of subordinate learning; (2) lack of retention of a subordinate skill because of insufficient practice; and (3) inadequate instruction to permit the integration of subtasks.

Number and Use of Subjects

Another interesting research question relates to the selection of sample subjects to be used as a data source for various formative evaluation efforts. There are advocates of the tutorial model involving formative evaluation data obtained from single learners in linear fashion with repeated try-outs. There are advocates of the large-group model involving formative evaluation of the prototypes with larger samples that are more generally representative of the intended target population and employing more sophisticated measurement and analysis techniques. Others have suggested the use of small-group try-out in conjunction with debriefing sessions.

A study by Robeck (1965) was designed to test the feasibility of the use of a single student as the data source for formative evaluation leading to the revision of an instructional program. An initial program revision was based upon the reaction of a sixth-grade boy, and the second revision was based upon another individual reaction. Three versions of the program (the prototype, revision 1, revision 2) were then administered to three matched groups of students. The performance on the two revised editions was significantly better than that on the prototype, although the performance on the second

revision was not found to be superior to the performance on the first revision. This study demonstrates that observation of a single student is an economical method for significantly improving instruction; its difficulty, however, lay in developing procedures for selecting "an appropriate student." The use of a single student was employed by other product developer/evaluators as well (Fleming, 1963; Markle, 1967) to mention but a few.

In a variation of the single-student procedure, Abedor (1972) used a small group combined with a debriefing session. Revisions based upon his formative evaluation procedure produced statistically significant outcome results over prototype material "on 9 out of 12 dependent measures in three separate field experiments." This is perhaps a not very startling result. As an interesting by-product of the study, the analysis of the data suggested that obtaining feedback through procedures utilizing a student debriefing session may serve as instructional product development training. The effect of such "training" may result in improved quality of subsequent prototypes directed at the same target audience; however, Abedor presents no data to support his supposition.

A question that persists and that is of continued interest to those engaged in the development and formative evaluation of instructional systems is the question of timing for the employment of the various data sources. How polished must a prototype version be before it is submitted to formative evaluation? This question has been raised by many writers, including Gilbert (1962), Markle (1967), and Rummler (1965) et al. In discussing formative evaluation of films, Hovland, Lumsdaine and Sheffield (1949) have noted:

> "The stage at which evaluative research would appear to have its greatest potentiality for product improvement is well prior to the completion of a film, when suggestions derived from research can readily be incorporated into the final film production." (P. 259).

Use of Formative Evaluation Data

Since revision of instructional materials based upon formative evaluation data rests at the heart of instructional development, there have been a series of research studies conducted by those engaged in instructional technology and development that examine the use of formative evaluation data. The explanation for most of these studies is that they attempt to justify the time needed and cost spent in revision. A study involving instructional television was conducted by Gropper and Lumsdaine (1961) in which, on the basis of criterion data on a set of instructional lessons, revisions were made. They report improved learner results from revised material when experimentally compared with the originals. Three studies using film-strip mediated instruction, VanderMeer and Montgomery (1964), VanderMeer and Thorne (1964), and VanderMeer (1964), supported the use of evaluation data to improve instructional effects. VanderMeer, Morrison and Smith (1965) later used achievement data as a basis for the revision of two films.

The above-cited studies, and others of a similar nature, confirm one's common sense expectation that finding out something about program effects will improve one's next efforts. That is, formative evaluation data have been shown to make a difference in product performance. However, there are a number of reservations regarding this research, based, in part, upon the state of development of the technology of instructional revision at the time of the studies. For one thing, the studies above are unalterably confounded, in that single individuals or teams used the formative evaluation data for revising the materials. Thus, unarticulated, unintended bias might have enhanced the revision versions. For a second thing, there was little evidence that the revisions made in these studies were controlled by the formative evaluation data (either the types of data collected or the pattern of responses obtained). The studies merely demonstrate that, after collecting some formative evaluation data, someone did something to improve a single instructional sequence. A serious issue in this research, then, is the lack of generalizability of the findings. How can we demonstrate that researchers other than the particular experimenters can obtain results favoring empirical methods? How can we demonstrate that it is the formative evaluation data that have impact upon improved program performance rather than some other variable, such as simply increased time to consider the material?

A particularly interesting line of research,, pertinent primarily to the question of generalizability, was established by Rosen in a 1968 study. Rosen employed the programmer as the experimental unit. Two groups of programmers were randomly assigned to experimental treatments. Both groups were given the same instructional materials and asked to improve them. One group was given formative evaluation data from tests of the program objectives given to a group of learners. The second group was given directions to use any criteria they wished to revise the program; they were not provided with evaluation data of any sort. The "improved" programs were then administered to randomly-assigned sets of learners. Student performance on both sets of revised materials was found to be superior to that achieved on the posttest of the original, unrevised, program. The performance of the program revised on the basis of formative evaluation data was found to be higher than that of the intuitively-revised programs.

Variations of the Rosen study have been conducted. Sulzen (1972) used programmers as the experimental unit in a study using five revisions, each of two different programs. Baker (1970) used seven different instructional programs in her study.

Each program was written by a different person and tried out on three learners. Programs were then randomly reassigned to writers along with formative evaluation data on the program they were to revise. After revisions were made, the programs were randomly assigned to subjects, following the model demonstrated by Rosen. Obtained performance differences favored the revised versions of the programs.

While the above studies provide gross information the impact of formative-evaluation data upon revision is still obscured. What does or should the program writer do with formative-evaluation data? In what form should formative-evaluation data be provided? What rules are there to guide the way in which developers might use such data?

The study by Baker (1970), previously cited, attempted to provide rules for developers to assist them in making use of formative-evaluation data. However, this study did not experimentally control for the use of these revision rules. A later study by Robinson (1972) trained programmers in the use of revision rules. Students who used revisions prepared by the trained revisers achieved a significantly higher mean than did students who used revisions prepared by untrained revisers. These studies both focus upon the training of product developers to use formative-evaluation data more effectively for revision purposes.

The question of the kind and level of formative-evaluation data that should be considered satisfactory for moving a potential instructional product to its next stage of development is one which has not been examined to any extent. Edmonston, Kunetka, and Newman (1973) performed a study in which they examined the question of whether or not evaluators and product developers agree as to when a product should be advanced from one developmental stage to the next and whether or not they rely upon the same sources of data for making that decision. In 21 simulated situations, descriptions were provided of a program in a specified stage of product development. Five product developers and five evaluators were asked to decide, on the basis of cost, performance, and spin-off data, whether or not the program should be allowed to progress to the next product developmental stage and which of the three indicated data sources were relied upon most and least in making this decision. Both groups rated the importance of three data sources similarly at each of the stages. Interestingly, evaluators seemed more willing to allow products to proceed to the next developmental stage at the early stages of development. While the study has too small a sample to provide significant empirical data, it does serve to remind us of the potential utilization of the more sophisticated JAN methodology (Christal, 1963) for "capturing the policy" upon which individuals base their decisions. The Christal use of regression techniques for this purpose might be most appropriate as a tool for research on formative evaluation relative to revision decisions.

In related research, Dick (1968) conducted a study in which he determined the functional state-of-the-art in revision. In essence, his study dealt with the question of how evaluation data were used by revisers and what kinds of evaluation data were used in making decisions. In his study of revisers' behavior, Dick found that revisers attended to reviewer comments and error rates within the program, and infrequently looked to performance on the posttest as a source of data for revisions.

Another line of research is represented by a study conducted by Bank (1972). The research issue is whether the collection of formative-evaluation data interferes with the learning of subjects. In the treatment conditions, subjects, who were classroom teachers engaged in an in-service training program, were requested to respond during the course of instruction to two-page questionnaires concerning the value and effectivensss of the product. The outcome measures for both treatment and control groups were a cognitive posttest and an attitude measure. Her findings suggested that learning was not depressed by the inclusion of questions relating to the clarity of presentation, suitability of examples, etc. Her findings were partly replicated by Baker and Quellmalz (1972).

FORMATIVE EVALUATION MODELS

In this section of the chapter we will examine various formative evaluation models and consider their appropriateness for use in the evaluation of instructional products. We will focus initially upon the formative-evaluation components of total models and upon formative-evaluation procedures that are derived from a broad conception of evaluation. Later in the chapter we will consider procedures which are built into instructional product-development models or which are derived primarily from an instructional product-development point of view.

Evaluation Models

Within the field of evaluation, if one will consider Scriven (1967), who coined the term "formative evaluation", as the father of formative evaluation, then surely Cronbach, who said "the greatest service evaluation can perform is to identify aspects of the course where revision is desirable.," (1963) must surely be thought of as the godfather. The efforts first of Cronbach and then of Scriven turned the eyes of the field of evaluation to a concern for evaluating programs in order to make revisions and attain improvement. Scriven (1972) and Stake (1972) imply that it is best that the formative evaluation be performed by an agent external to the program. Scriven calls his approach "goal-free evaluation" and establishes a rule in which the formative evaluator collects her or his own data and renders his assessment of the actual effects of the project. The evaluator operates without knowledge of the claims made for

the program by its developers. Thus, no statement of purposes, goals, or objectives are admissible data for Scriven. The evaluation proceeds without coercion of documents. The counterpart of "goal force evaluation" is called "responsive evaluation" by Stake. Like the goal-free evaluator, the responsive evaluator is primarily external to a program or product-development activity, and therefore has a certain independence and objectivity that is presumed not to be present in an internal evaluator. The formative-evaluation role portrayed by both Scriven and Stake is in many respects analogous to that of a financial auditor who enters the scene, makes observations based to some extent upon guidelines but heavily upon her or his own choosing, and, by making evaluation observations kown, manages to assure the stockholders that either all is well, or that improvements are underway, or, at minimum, that some irrefutable source is keeping the system honest.

Extrapolating from the writings of Stufflebeam (1971) and of Alkin (1969), one notes that their views of formative evaluation as it relates to instructional-product development are less dependent upon an external evaluator, primarily because they emphasize the necessity of structuring evaluation so that it "serves" decision-making purposes by producing appropriate and timely information. Thus, the possibility exists in these models that the evaluator can be a part of the development team, if that arrangement seems most efficacious for providing the kinds of information that are required for the decisions that must be made. Alkin (1973) has expressed his concern for evaluation being integrally tied to the specific design-purpose by arguing for the necessity of the evaluation role being "framed by the decision context."

The total evaluation model developed by Provus (1971) offers another set of guidelines for the formative evaluation of instructional products. Provus' total-evaluation model was designed from the vantage point of the development of instructional programs and materials within an individual school district. Thus, while it also is not focused specifically upon evaluation related to instructional product development, the model has considerable relevance and offers insight into the appropriate procedures for the formative evaluation of instructional products.

What must be concluded from this examination is that while the term "formative evaluation" was invented by Scriven and pressed to the collective bosom of the evaluation community, these conceptions of evaluation have for the most part been addressed to the evaluation of total educational programs with specific context designations, rather than the evaluation of instructional products being devised by developers. In fact, even before the entrance of the evaluation community into the field of formative evaluation, there existed an evaluation process integrally tied to and a part of instructional product development, particularly as it related to programmed learning, even though it had not been so elegantly named and identified.

Instructional Product Development Technology

Now let us consider the formative evaluation models and procedures that grew from the instructional product-development movement. Instructional product-development as a technology emerged from the programmed instruction and the curriculum development movements of the late nineteen-fifties and early nineteen-sixties. By the beginning of the nineteen-seventies, the field had achieved a dramatic growth-rate. One evidence of the growth of the field during the past decade is demonstrated by the evolution of over 20 regional laboratories and R&D centers, many of which have been leaders in the development of formatively-evaluated instructional products. Furthermore, there have recently been many publications in the literature describing various schemes and models of product-development processes (Baker, 1973; Banathy, 1968, Borg and Hood, 1968; Briggs, 1970; Glaser, 1966; Mager and Beach, 1967; Stowe, 1969; Schutz, 1970; Sullivan, 1971; Locatis and Smith, 1972; Johnson, 1972).

The complexity of the educational product-development models varies considerably, from conceptual schemes (Schutz, 1970) to comprehensive, step-by-step procedural checklists (e.g., Borg and Hood, 1968). But, at the core of each model, regardless of its complexity or simplicity, regardless of its esotericism or practicality, is the realization that product development and formative evaluation are intertwined as snake and staff and that product revision depends upon the generation of formative-evalution data.

Numerous examples can be cited of specific applications of product development-based evaluation models. For instance, in a description of the procedures used in improving Project PLAN materials, Rahmlow (1969) describes the use of field data from 13 school districts. While this was not a research study, it is presented in this section in order to "balance" the discussion on data sources, and to provide the reader with a description of the categories of data commonly used in large-group formative evaluation. The system reported that Rahmlow (1969) relied on the following types of "hard" and "soft" data:

Hard data included the following categories:

1. item statistics of module tests indicating the number of students selecting each option
2. frequency distribution of student scores on module tests on an objective-by-objective basis
3. frequency distribution of student scores on module tests on objective-by-objective by Teaching-Learning UNIT (T.L.U.) basis
4. cross-validation data on parallel items administered at various time intervals

Soft data included observational and teacher-observed behavior:

1. questions asked of the teacher about T.L.U. used in conjunction with item analysis provide clues for improvement. questions cover overall evaluation of T.L.U., learning activities, and tests
2. student responses to questions (concerning amount of time, assistance, attitudes, etc.) are also used in conjunction with hard data
3. project personnel working on improvement make periodic visits to Project PLAN Schools to maintain a realistic context for data interpretation

Centers and Laboratories have also produced many documents illustrating formative-evaluation processes. (See Lindvall and Cox, 1970; Light and Reynolds, 1972; King, 1970; Nimnicht, 1970; Niedermeyer and Ellis, 1970; and Miller, 1972.) There are examples that might be provided of applications of formative-evaluation procedures to product development. A search of the ERIC system reveals the following listing: Grobman, 1968; Bogatz and Kurfman, 1966; Branson, 1971; and Raynor, 1972.

There is no question but that formative-evaluation procedures have been most thoroughly refined and implemented at the regional laboratories and R&D centers involved in developing instructional products. There is also substantial evidence of this use by other agencies or individuals engaged in the development of instructional products. Some have argued, however, that the specificity of the formative-evaluation procedures which to a large extent were developed at these laboratories and centers have provided only "techniques applicable to simple stimulus configurations such as textual programmed instruction" (Abedor, 1972). Abedor suggests that try-out and revision procedures appropriate for programmed instructional materials are not appropriate for multimedia systems. We would question that suggestion and point out that many of the formative-evaluation procedures employed at these laboratories and centers indeed have been applied to multimedia systems (Lindvall and Cox, 1970; Hood and Johnson, 1969).

We maintain that the evidence is clear that well-applied formative-evaluation procedures are equally appropriate for instructional products of all types. Indeed, the formative-evaluation procedure proposed by Sedlik (1971) for evaluating instructional films and television productions is not significantly different from other formative-evaluation models. Diagnostic pretesting using a "feedback-oriented evaluation context by applying systems techniques" (Sedlik, 1971) is the essence of all formative-evaluation procedures.

Other formative-evaluation procedures have been developed which are related specifically to mediated instruction, in particular, films. (See Finn, 1956; Kaufman, 1964.) Even though the overwhelming number of films have not been subjected to formative try-out, there are numerous examples of individual applications of the use of formative-evaluation procedures in the development of mediated instruction (see Sedlik, 1971).

In the next section of this chapter we will describe in detail the steps involved in performing a formative evaluation of an instructional product.

PROCEDURES FOR FORMATIVE EVALUATION

Descriptions or recommendations of procedures for conducting the formative evaluation of instructional products must always be muted in their prescriptiveness. Clearly, different projects will require different distributions of attention, and, certainly, there is no "right" way to conduct an evaluation. On the other hand, there may be some wrong ways. A well-meaning evaluator might be able to improve her or his contribution by the use of procedures which have been tried successfully by others. The purpose of this section is to identify some critical elements in the conduct of formative evaluation of instructional products during development. The procedures described here should be viewed as points of departure. More importantly, a set of guiding questions will be described which should assist the evaluator in determining whether the implementation of formative assessment has been circumspect, complete and economical.

Data and Decision-Making

"Data-for-decision-making" is the basis for much formative evaluation; perhaps, in fact, it is the slogan of the activity. However, the proper implementation of this concept is not necessarily obvious. In order to gather data to assist in decision-making, the evaluator must have a good idea of what the possible decisions may be at any stage of development. With such knowledge, the evaluator is able to procure data that apply to these decisions. Therefore, data for decision-making may imply not only that an overall assessment of a program is required, but that every type of data collected will contribute to the decision-making process itself. If one pursues the implication that all data collected should be reflected in decision processes, one is faced with an immediate concern; the range of possible decisions should be articulated in advance of the plan for data collection.

A more typical procedure may involve the development of an innovation and its subsequent trial, to see if it "flies," "waves," or perhaps "bombs," in the metaphor of the vernacular. The evaluator may be doing a through assessment of the program effects, yet in a formative context, the evaluator must also collect data which will permit the expedient revision of the program so that improvement in results can be anticipated in future trials. Thus, the sorts of information tapped and analyzed must be of such a nature as to assist in the revision process.

Suppose, for simplicity's sake, one were involved in the development of a program that had a single objective. Many evaluators would collect data on post-instructional performance with respect to that objective, provide the achievement level to the development staff, and thereby conclude her or his effort. Such data are likely to be sufficient when the program is successful. More finely-grained information would not be necessary, for program in which revisions are not required. But most projects do not develop successful innovations on the first trial. To be of service, the formative evaluator must understand the range of options that the project has available as s/he anticipates the worst possible outcome from evaluation; what should be done if significant components of the program fail? The kind of answer one gives to this question is clearly a function of the stage of development of the program at the time of data collection.

Two distinct development stages have been identified by writers in the field (Markle, 1966; Schutz, 1970). The first stage can be termed the initial development stage, a phase of activity that culminates in a prototype instructional sequence for testing. The second phase of development occurs following the successful test of the prototype, when additional components and procedures are added in an attempt to make the program useful in a real setting. Each of these development stages has an associated set of formative-evaluation activities.

The purpose of evaluation during the prototype stage is to verify whether the instructional product is capable of producing the desired outcomes. Data collected to assist in this judgment must anticipate the possible range of decision in the prototype-development phase. Data gathered at the operational stage of development should pertain not only to the level of learning achieved by the students, but also to the degree to which the program functions as anticipated and to which it receives approval by administrators and teachers. There would seem to be a close relationship between the degree to which the program is likely to be an instructional success (that is, meets its objectives) and the practicality-of-use dimension.

There are three general areas of concern in formative evaluation of instructional development. The question of instructional design should be paramount during the earlier phases of instructional development, the prototype stage. The evaluator turns her or his interest to an additional set of concerns during the operational development stage, in attempts to determine the usability of the program in the range of intended settings. A third element for the attention of the evaluator is the process and organization of the development institution itself. This question is of particular concern for agencies with a lengthy anticipated life-span. Each of the areas--instruction, operations, and development processes--will be briefly explored in light of the data it is desirable to collect at each stage.

Initial Development: Instructional Sequence

The desired outcome of the initial development effort is the production of a prototype or prototypes which can be tested to determine their power in assisting the learner to meet program goals. In the initial development cycle, the designer makes her or his best attempt to generate an effective sequence. Presumably, careful analysis enters into the choices of the instructional **principles** for use, the instructional **approaches** selected, and the adequacy of the content in terms of comprehensiveness, importance, and accuracy. An assumption made here is that some sort of review of goals and tryout of measures have preceded the instructional design phase, but clearly, formative evaluators must verify this presupposition.

What sort of data would assist in the potential revision of the instructional prototype? Two major classes of formative-evaluation information are necessary: (1) Information about the effects of the program, and (2) information that should assist in the remedy of program deficiencies.

Under category one the primary question at the prototype phase relates to the success of the program in meeting its goals. The relevant data source for this question is the student performance on the criterion set of tasks. Clearly pretest data will be necessary to determine the magnitude of improvement evidenced by the students. A second concern in terms of program effects could be the reaction of the learner to the objectives, instructional sequence, and approach of the instruction, particularly for these programs in which learner affect toward the program is critical. Notice that questions regarding the learners' interest in the format selected or media used is not critical at this phase of development, for prototype tests often employ rough-cut versions of film, storyboards rather than fully developed renditions, or ditto-mediated rather than three-color, type-set, versions of text. A third area of concern, even at the prototype-development phase might be producing. The conduct of free-discussion interviews with subjects may help to uncover types of misinformation that the product is unwittingly generating. In summary, the three types of data important to determining product effectiveness in the prototype or initial development stage are: (1) Student performance of post-instructional tasks, (2) Student attitudes and satisfaction with program, and (3) Unanticipated side effects engendered by the program.

The second category of data in the initial development tryout relates to providing diagnostic information to the developer in order that the product may be altered if necessary. Without such a data source, each modification will represent merely another unsubstantiated hypothesis, and the probability of success will not increase with each formative trial. Each revision would be, rather, an independent event, with the probability that the revision will improve the situation unchanging from

revision to revision. Such a situation would clearly not be desirable in development settings where schedules of completion and funding are often contingent upon the demonstration of program success in a planned period of time.

Data sources to assist in the revision of initial tryouts must be pertinent to the types of modifications that the developer may make. The following data issues are presented for consideration.

Expert Judgment

Expert judgment regarding instructional products probably should be secured when the specifications of the product are being developed. Experts can respond to the selection of goals and content as well as to the adequacy of measures. However, one might also wish to subject prototypes to the scrutiny of content specialists who, for instance, might detect subtle errors in subject-matter or sequence.

Program-Response Data

What is the pattern of errors that the learner makes on practice exercises in the instructional program? What other sorts of records of student behavior can cue the developer to the particular sections of the program that are not assisting learning? Data may be assembled from the actual responses the student makes in the process of instruction.

Program-Critique Data

How helpful does the learner find particular examples which have been provided in the instruction. Is the organization of the materials suitable? Is the level of language appropriate? Are there sections that provide particular difficulty or confusion? Does the approach make sense to the learner? Data for use in answering these questions, all of a diagnostic nature, may be obtained in at least three ways. First, learners may be instructed to cricle difficult words, write comments, or place question marks in the materials that they are exposed to. If printed materials are not utilized, then periodic, brief formative checks might be made at regular intervals, in which the learner is asked to provide answers to questions similar to those listed above. Or single learners may be queried as they proceed through the program. A third source is a post-instructional diagnostic questionnaire, which asks students to give their analyses of and suggestions for the revision of the sequence. An interview can also provide informatin that should enable the developer to identify particular aspects requiring revision.

Procedural Considerations in Data Collection

It should be clear that the initial development tryout represents a case in which the formative evaluator wants very detailed, very specific information. The formative evaluator must plan to gather all of the data which possibly might be useful in making the revision at this point, particularly due to her or his low level of confidence that the program will succeed. The developers' design has been a carefully executed hypothesis, but no one has reason to believe that it will be successful. Thus, the formative evaluator should consider two points in the planning of the initial development evaluation: first, that relatively few students should be exposed to untested instruction; second, that students selected for the field trial should be expected to provide both post-instructional and performance data. The rich variety of data expected requires careful analyses, analyses that may often be resistent to easy summary. These two factors, exposure to untried instruction and the need for rich data, combine to suggest that the initial development tests be conducted upon a relatively small number of students and that the data collection procedure be designed to provide the completeness of data required. Smaller samples represent a loss in the reliability of the data and a shortening of turn-around time for analysis and revision. Since initial development often has an exploratory character, it would be unwise to interrupt the development process with lengthy intermissions caused by the analysis of too much data. Thus, the use of smaller samples, more intensively analyzed, would appear to be appropriate at this stage of development.

Operational Development: Practicality

Following the occurrence of a successful initial development try-out, an event which is likely to require more than one cycle of development-trial, the development operation shifts focus to a concern with the usability of the program in the range of anticipated settings. The attention of the formative evaluator conducting a field trial must similarly be directed to issues of practicality, providing, of course, that the program still produces the desired student performances. Recalling the classes of data identified in the initial development tryout, we find that (1) the effects of the program should be assessed; and (2) diagnostic information should be obtained which might assist in revision. Again, what are the questions of concern to the developer at this point in the project? First, of course, is the extent to which the program is succeeding with the learners. The same questions regarding program effects in terms of learner achievement which were identified in the initial tryout stage must be reconsidered:

1. Student performance on post-instructional tasks
2. Student attitudes and satisfaction with the program
3. Unanticipated effects engendered by the program

Such data are collected so as to ascertain whether

or not the translation of the program into operational status has in some way reduced the effectiveness of the learning sequence.

There are several other questions which must be addressed at this stage that require diagnostic data. If a proper initial development try-out had been conducted, few questions should remain regarding the instructional sequence itself. The revision questions of interest will therefore revolve around the extent to which the user is satisfied with the materials, the administrative feasibility with which they can be used, and the students' responses to the integration of program components and the format chosen for instruction.

Within the category of data measuring program effects, additional information may be sought about the following concerns:

1. Student satisfaction regarding the format and the integration of the program components.
2. Administrator or instructor ability to implement the program.
3. Administrator or instructor satisfaction with the program.

Student satisfaction can easily be tapped on a questionnaire, administered at the conclusion of the program. Administrator ability to implement the program can be assessed in more than one way. Observers may record teacher behavior involving critical components of the program. Teachers themselves may prepare reports on their daily successes and failures in using the materials. Query-telephone lines might be made available, so that administrators can call and ask pressing questions about program use. A record of the sorts of questions asked may be useful. Teachers' satisfaction with the program may be assessed through both questionnaire and interview procedures.

As has been noted, the second major class of data in an operational field-test relates to the provision of diagnostic information necessary for program revision. Since the emphasis in a field test is different from the emphasis in the initial development test, the potential kinds of revision activity differ as well. Rather than concern with learning principles, adequacy of examples, and viability of the approach, revision will more likely be concerned with clarifying and simplifying directions to students and teachers, so that the program can work within the limitations imposed by the instructional and administrative settings. Questions regarding the integration of materials, such as whether or not materials were actually employed in the sequence anticipated, should be answered. Information regarding the feasibility of the selected presentation media should be gathered (for instance, if hardware was required, were there frequent technical problems, e.g., television monitors that were fuzzy, projectors that didn't operate?). Rather than research-oriented learning principles, the developer will utilize knowledge about management in solving whatever problems that the field trial uncovers.

The field trial has another important function, in that it serves to identify (or to refine) the teacher/user training requirements to which the program should attend. If teachers are systematically misusing given components and if the developers wish to stay with their original design because of extremely positive prototype data, then they must add a teacher-training sequence to assure that intended procedures will be followed.

Clearly because several users are required, the number of students used in an operational field-test is larger than in prototype tests. However, note that the difficult data, observation, interview, and daily records, ae generated by the teacher or program administrator rather than by the student. Again, there should be concern to avoid an over-accumulation of data. Student data, where large numbers of students are utilized, should consist of relatively easily-quantifiable information. Any more-intensive scrutiny should probably be handled on a sampling basis, stratifying, of course, by teacher.

Evaluation of Process: Staff Procedures

Formative evaluation necessarily attends to the outputs of a given project (how well the progam works, what probably needs changing), but that is not all that it attends to. The formative evaluator may also be concerend with the evaluation of the development process--how the development staff is proceeding with its business. If a formative evaluator determines that a consistent problem occurs in the program, for instance, that the texts contain many typographical errors, then s/he may suggest that certain development activities be modified in order to avoid the recurrence of the problem. Formative evaluation can identify problems in communication, for instance, in cases in which there are apparently overlapping chains of authority that confuse staff. Formative evaluation may also identify redundancies in or areas of omission from the developmental process.

The impetus for the evaluation of staff processes derives from organizational research, and subsequent recommendations for revision activities will generally be based upon knowledge of social interactions and behavioral science, rather than of learning psychology.

Certainly, no formative evaluation of an organizational process in an agency should be conducted unless that agency will be functioning for some time. A short-term project may need to have certain glaring errors identified but the careful analysis of interaction patterns would probably not be useful. By the time the analysis was implemented, the project may have been completed.

SUMMARY

In this chapter we have tried to describe the research pertinent to the activity of formative evaluation. While the number of good examples of

formative evaluation is expanding, research into the process is relatively limited. Perhaps a compromise to the difficult task of accumulating research data on formative evaluation might be suggested. When formative evaluation activities have been successful in terms of program effects and staff satisfaction, then detailed technical reports might be made available to the evaluation public. Formative evaluation might improve as a consequence of the technology developed in the course of finding solutions to development data problems. Both the increase in technology as a result of dissemination of successful procedures and the pursuit of experimental research on the process of evaluation might ultimately remove formative evaluation recommendations from the realm of seers and clairvoyants and bring it near, if not into, the domain of scientific application.

REFERENCES

Abedor, A.J. Second draft technology: Development and field test of a model for formative evaluation of self-instructional multi-media learning systems. Viewpoints, 1972, 48(4), 9-44.

Alkin, M.C. Evaluation theory development. Evaluation Comment, 1969, 2(1), 2-7.

Alkin, M.C. Evaluation and decision making. Planning and Changing, 1973, 3(4), 1-4.

Anderson, R.D. and Faust, G.W. The effects of strong formal prompts in programmed instruction. American Educational Research Journal, 1967, 4(4), 345-352.

Baker, E.L. The technology of instructional development. In R.M.W. Travers (Ed.), Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973.

Baker, E.L. Generalizability of rules for empirical revision. AV Communication Review, 1970, 18(3), 300-305.

Baker E. and Quellmalz, E. Research-Based Techniques for Instructional Design. Los Angeles: National Center for Educational Research and Development, 1972.

Baker, R.L., Schulz, R.E., and Sullivan, H.J. Internal control procedures in the development of instructional materials. Technical Reports: Research on Procedures for the Development of Instructional Materials. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratories, Aerospace Medical Division, 1967.

Banathy, B.H. Instructional Systems. Palo Alto, CA: Fearon Publishers, 1968.

Bank, A. The effect of evaluating instructional products on user test performance, attitude, and revision judgment. Unpublished doctoral dissertation, University of California, Los Angeles, 1972.

Bogatz, G. and Kurfman, D. Intra-urban unit: ETS evaluation report, limited field trials high school geography project. Princeton, N.J.: Educational Testing Service, 1966.

Borg, W.R. and Hood, P. The twenty-seven steps in the development program. Berkeley, CA: Far West Laboratory for Educational Research and Development, 1968.

Branson, R.K. Formative evaluation procedures used in designing a multimedia physics course. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Briggs, L. Handbook of Procedures for the Design of Instruction. Pittsburgh: American Institutes for Research, 1970.

Christal, R.D. JAN: A Technique for Analyzing Individual and Group Judgment. Lackland Air Force Base, Texas 6570th Personnel Research Laboratory, Aerospace Medical Division, February 1963 (PRL-TDR-63-12).

Cronbach, L. Evaluation for course improvement. Teachers College Record, 1963, 64(8), 672-683.

Crowder, N.A. Automatic tutoring by intrinsic programming. In A.A. Lumsdaine and R. Glaser (Eds.), Teaching Machines and Programmed Learning: A Source Book. Washington: National Education Association, 1960.

Dick, W. A methodology for the formative evaluation of instructional materials. Journal of Educational Measurement, 1968, 5(2), 99-102.

Edmonston, L.P., Kunetka, J.W., and Newman, M.A. Undertaking program comparisons in curriculum evaluation. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

Finn, J.D. The testing of public relations audiovisual materials. PR: The Quarterly Review of Public Relations, 1(10), 1956.

Fleming, M. What is a good film? Journal of the University Film Producers Association, 1963, 15, 8-9, 18.

Gagne, R.M. and Paradise, N.E. Abilities and learning sets in knowledge acquisition. Psychological Monographs: General and Applied, 1961, 75(14), 1-23.

Gilbert, T.F. Mathetics: The technology of education. Journal of Mathematics, 1962, 1(1), 7-73.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18(8), 519-521.

Glaser, R. Psychological bases for instructional design. AV Communication Review, 1966, 14(4), 433-449.

Grobman, H. Evaluation Activities of Curriculum Projects: A Starting Point. AERA Monograph Series on Curriculum Evaluation, No. 2. Chicago: Rand McNally, 1968.

Gropeer, G.L. and Lumsdaine, A.A. The Use of Student Response to Improve Televised

Instruction: An Overview. Studies in televised instruction. Pittsburgh: American Institutes for Research, 1961.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., and Lundin, S. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study From the MINNEMAST Project. CSE Monograph Series in Evaluation, No. 1, Los Angeles: Center for the Study of Evaluation, 1973.

Hively, W., Patterson, H.L., and Page, S.H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5(4), 275-290.

Hood, P.D. and Johnson, J.N. The development and evaluation of television workshops in human relations. Berkeley, CA: Far West Laboratory for Educational Research and Development, April 1969.

Hovland, C.I., Lumsdaine, A.A. and Sheffield, F.D. Experiments on Mass Communication. Princeton, N.J.: Princeton University Press, 1949.

Husek, T.R. and Sirotnik, K. Matrix sampling. Evaluation Comment, 1968, 1(3), 1-4.

Johnson, T.J. A macro view of process evaluation: Evaluation in the context of product and program development in laboratory and research and development centers. Viewpoints, 1972, 48(4), 79-110.

Kaufman, R.A. The system approach to programming. In G.D. Ofiesh and W.C. Mierhenry (Eds.), Trends in Programmed Instruction. Washington: Department of Audiovisual Instruction, National Education Association and National Society for Programmed Instruction, 1964.

King, I.L. A formative development of a unit on proof for use in the elementary school (Parts 1, 2, 3). Madison, Wisc.: Wisconson University Research and Development Center for Cognitive Learning, January 1970.

Komoski, P.K. The editor's guide to this issue. (Editoral) Programmed Instruction, 1956, 5(1).

Komoski, P.K. Statement to the U.S. House of Representatives Committee on Education and Labor. In Hearings: To Establish a National Institute of Education, 92nd Congress, First Session, Washington, U.S. Government Printing Office, 1971.

Light, J.A. and Reynolds, L.J. Debugging product and testing errors: Procedures for the formative evaluation of an individualized mathematics curriculum. Viewpoints, 1972, 48(4), 45-78.

Lindvall, C.M. and Cox, R.C. Evaluation as a Tool in Curriculum Development: The IPI Evaluation Program. AERA Monograph Series on Curriculum Evaluation, No. 5, Chicago: Rand McNally, 1970.

Locatis, C. and Smith, F. Guidelines for developing instructional products. Educational Technology, 1972, 12(4), 54-57.

Mager, R.F. Preparing Instructional Objectives. Palo Alto, CA: Fearon Publishers, 1962.

Mager, R.F. and Beach, K.M. Developing Vocational Instruction. Palo Alto, CA: Fearon Publishers, 1967.

Markle, S.M. Empirical testing of programs. In P.C. Lange (Ed.), Programmed Instruction. The Sixty-sixth Yearbook of the National Society for the Study of Education, Part II. Chicago: NSSE, 1967.

Miller, G.R. The use of formative evaluation procedures in the development of a mathematics laboratory. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43(2), 205-216.

Niedermeyer, F.C. and Ellis, P.A. The Development of a Tutorial Program for Kindergarten Reading Instruction. Inglewood, CA: Southwest Regional Laboratory, 1970.

Nimnicht, G.P. A revision of the basic program plan of education at age three. Berkeley, CA: Far West Laboratory for Educational Research and Development, August 1970.

Provus, M. Discrepancy Evaluation for Educational Program Improvement and Assessment. Berkeley, CA: McCutchan Publishing Corp., 1971.

Rahmlow, H.F. Use of student performance data for improvement of individualized instructional materials. In Proceedings of a Symposium. Washington: American Psychological Association, 1969.

Raynor, G.T. An empirical study of a methodology for the revision of systematically designed educational materials. Tallahassee, FL: Computer-Assisted Instruction Center, March 1972.

Robeck, M.D. A study of the revision process in programmed instruction. Unpublished master's thesis, University of California, Los Angeles, 1965.

Robinson, T.J. Replicable training in revision techniques. Unpublished doctoral dissertation, University of California, Los Angeles, 1972.

Rosen, M.J. An Experimental Design for Comparing the Effects of Instructional Media Programming Procedures: Subjective VS. Objective Revision Procedures. Final report. Palo Alto, CA: American Institutes for Research in Behavioral Sciences, 1968.

Rummler, G.A., Brethower, D.M., Markle, D.G., Shrader, A.W., and Smith, D.E.P. Programmed Learning: A Practicum. Ann Arbor, MI: Ann Arbor Publishers, 1965.

Schutz, R.E. The nature of educational development. Journal of Research and Development in Education, 1970, 3(2), 39-64.

Scriven, M. The Methodology of Evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally,

1967.

Scriven, M. Pros and cons about goal-free evaluation. Evaluation Comment, 1972, 3(4), 1-4.

Sedlik, J.M. Systems Techniques for Pretesting Mediated Instructional Materials. Los Angeles: Education and Training Consultants, 1971.

Shoemaker, D.M. Evaluating the effectiveness of competing instructional programs. Educational Researcher, 1972, 1(5), 5-8, 12.

Silberman, H., Coulson, J., Melaragno, R., and Newmark, G. Use of Exploratory Research and Individual Tutoring Techniques for the Development of Programming Methods and Theory. Final report, NDEA Project 7-14-000-181. Santa Monica, CA: Systems Development Corporation, 1964.

Stake, R.E. The countenance of educational evaluation. Teachers College Record, 1967, 68(7), 523-540.

Stake, R.E. Responsive evaluation. Unpublished manuscript, University of Illinois, 1972.

Stowe, R.A. Case Studies in Instructional Development. Bloomington, Indiana: Laboratory for Educational Development, Indiana University, 1969.

Stufflebeam, D.L., Foley, W.J., Gephart, W.J., Guba, E.G., Hammond, R.L., Merriman, H.O., and Provus, M.M. Educational Evaluation and Decision Making. Itasca, IL: F.E. Peacock Publishers, Inc., 1971.

Sullivan, H.J. Developing effective objectives-based instruction. Educational Technology, 1971, 11(7), 55-57.

Sulzen, R.H. The effects of empirical revision and the presentation of specific objectives to learners prior to programmed instruction upon the criterion behavior of military subjects. Unpublished doctoral dissertation, University of California, Los Angeles, 1972.

VanderMeer, A.W. An Investigation of the Improvement of Educational Filmstrips and the Derivation of Principles Relating to the Effectiveness of These Media. Study No. 2. University Park, PA: Pennsylvania State University, 1964.

VanderMeer, A.W. and Montgomery, R.E. An Investigation of the Improvement of Educational Filmstrips and a Derivation of Principles Relating to the Effectiveness of These Media: Study No. 3. University Park, PA: Pennsylvania State University, 1964.

VanderMeer, A.W. and Thorne, H.E. An Investigation of the Improvement of Educational Filmstrips and a Derivation of Principles Relating to the Effectiveness of These Media: Study No. 1. University Park, PA: Pennsylvania State University, 1964.

VanderMeer, A.W., Morrison, J., and Smith, P. An Investigation of the Improvement of Educational Motion Pictures and a Derivation of Principles Relating to the Effectiveness of These Media. University Park, PA: Pennsylvania, 1965.